Using *knitr* and LaTeX for literate lab notes

Boris Veytsman

TUG'2022

A cautionary story

How a student was frustrated trying to repeat research…

Am I sure *I* can understand my own work in ten years?

How do our colleagues in biomedical and experimental research solve this problem, when billions depend on it?

Laboratory notebooks

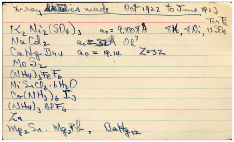
A Lab Notebook Is…

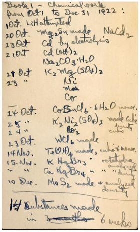
- *Complete record of procedures, reagents, data, and thoughts to pass on to other researchers*
- *Explanation of why experiments were initiated, how they were performed, and the results*
- *Legal document to prove patents and defend your data against accusations of fraud*

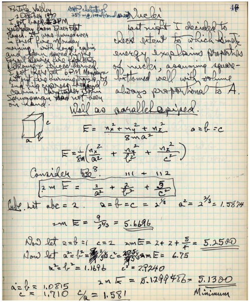
Philip Ryan (2012). Keeping a Lab Notebook. National Institutes of Health, Office of Intramural Training and Education. URL: https://www.training.nih.gov/assets/Lab_Notebook_508_(new).pdf

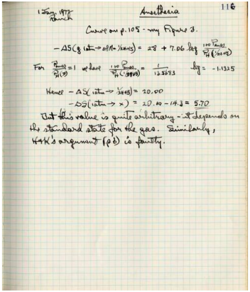
A classic example: Linus Pauling's notebooks

72 years of work (from 1922 to 1994): `http://scarc.library.oregonstate.edu/coll/pauling/rnb/index.html`









Lab notes as literate science

Knuth's insight: Your code is for computer. Your prose is for humans. ⇒ Literate programming[1].

Research situation: A paper (preprint, presentation) is just an *advertisement* of the research, but not the research. Research is a reproducible *environment* which includes computation and publication[2]. ⇒ Lab notes as literate science

[1] Donald E. Knuth (1992). *Literate Programming*. CSLI Lecture Notes 27. California: Stanford.

[2] Jill P. Mesirov (2010). "Accessible Reproducible Research". In: *Science* 327.5964, pp. 415–416. ISSN: 0036-8075. DOI: `10.1126/science.1179653`. URL: `http://science.sciencemag.org/content/327/5964/415`.

How do we keep lab notes?

The classic way: bunches of physical notebooks

- Very versatile: you can put there anything! *But*
- You cannot search efficiently (where is my `grep`?)
- Too many dead trees.
- Not too easy to keep after a couple of decades.

The modern way: electronic records

- Can be indexed, searched, compact! *But*
- Can we make them as versatile as physical ones?
- Can we make writing them as fast as scribbling?

What is in my lab notes? (1)

- Thoughts and ideas:

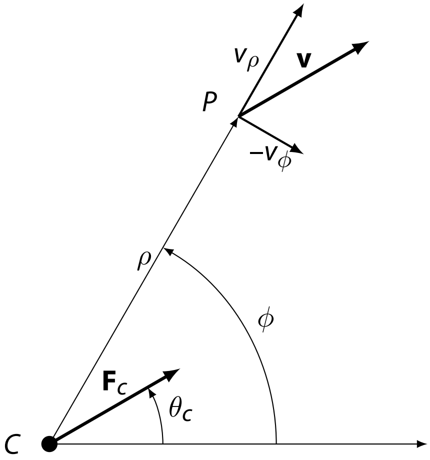
  It seems that cell diffusion inside a tissue is quite different if a different matrix around the tissue was used. This fact is quite inexplicable from the conventional picture of diffusion borrowed from the molecular physics. Indeed, how would a molecule inside a vessel "know" what is the vessel made of? One expects the measured diffusion not depend on the walls around the molecules.

- Equations

$$\mathfrak{a} = -\frac{l(l+1)c}{r},$$
$$\mathfrak{b} = -\frac{dc}{dr} - \frac{c}{r},$$
$$\mathfrak{c} = -\frac{a}{r} + \frac{db}{dr} + \frac{b}{r}.$$

What is in my lab notes? (2)

- Sketches



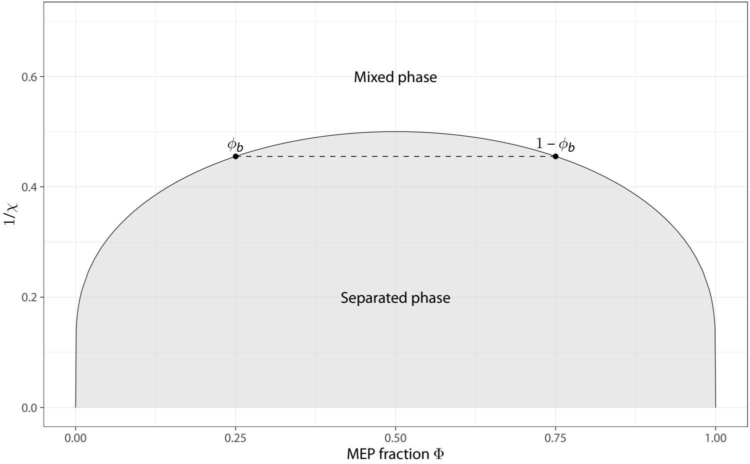
What is in my lab notes? (3)

- Program snippets…

```
chiPhi <- tibble(phi=c(seq(0,0.01, by=0.001),
                       seq(0.01,0.99,by=0.01),
                       seq(0.99,1, by=0.001))) %>%
mutate(chi= 1/(1-2*phi)*log((1-phi)/phi)) %>%
filter(!is.nan(chi))
chiPhib <- chiPhi %>% filter(phi==0.25 | phi==0.75) %>%
    mutate(label=c('$\\phi_b$', '$1-\\phi_b$'))
ggplot(chiPhi) + geom_line(aes(phi, 1/chi)) +
geom_polygon(data=chiPhi %>% add_row(phi=c(0,1), chi=c(Inf,Inf)),
             aes(phi,1/chi), fill='lightgray', alpha=0.5) +
    ylim(0,.7) + xlab("MEP fraction $\\Phi$") + ylab("$1/\\chi$") +
    annotate("text", x=0.5, y=0.6, label="Mixed phase") +
    annotate("text", x=0.5, y=0.2, label="Separated phase") +
    geom_point(data=chiPhib, aes(phi, 1/chi)) +
    geom_line(data=chiPhib, aes(phi, 1/chi), linetype='dashed') +
    geom_text(data=chiPhib, aes(x=phi, y=1/chi, label=label),
              nudge_y=0.025)
```

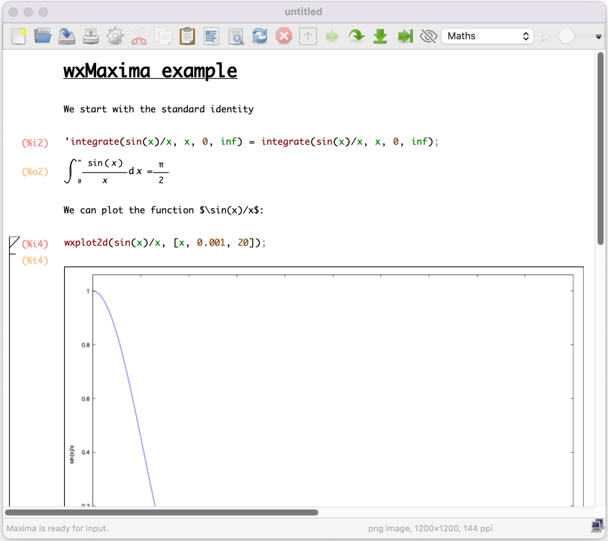
What is in my lab notes? (4)

- …and their results



Example: Computer algebra notebooks (1)

Many (all) commercial systems have them. Here is a free wxMaxima (`https://wxmaxima-developers.github.io/wxmaxima/`)



Example: Computer algebra notebooks (2)

wxMaxima example

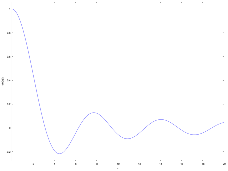
We start with the standard identity

(% i2) 'integrate(sin(x)/x, x, 0, inf) = integrate(sin(x)/x, x, 0, inf);

$$\int_0^{\infty} \frac{\sin(x)}{x} dx = \frac{\pi}{2} \qquad (\%\ o2)$$

We can plot the function $\sin(x)/x$:

(% i4) wxplot2d(sin(x)/x, [x, 0.001, 20]);



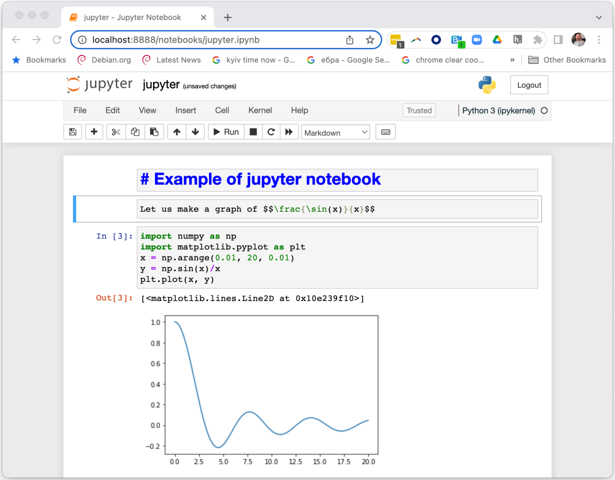
(% t4)

(% o4)

1

Example: Jupyter notebooks (1)

The great Jupyter project (`https://jupyter.org/`)



Example: Jupyter notebooks (2)

jupyter

July 17, 2022

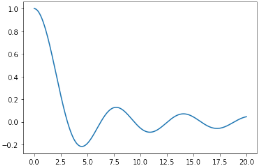
1 Example of jupyter notebook

Let us make a graph of

$$\frac{\sin(x)}{x}$$

[3]:
```
import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0.01, 20, 0.01)
y = np.sin(x)/x
plt.plot(x, y)
```

[3]: [<matplotlib.lines.Line2D at 0x10e239f10>]



[ ]:

1

My (humble) opinions about the examples

wxMaxima: Good for documenting equation manipulations. Not much convenient for everything else.

Jupyter: Good interface, especially when you play with code. Can incorporate many languages other than Python. *But*:

- Only a subset of LaTeX implemented. No label-ref, bibliography, etc.
- No support for sketches other than plots.

Common feature: LaTeX backend. Why not use LaTeX from the beginning?

My setup

Ideas:

1. I need the features of LaTeX: bibliographies, numbering, etc.
2. A bunch of `tex` files is easily searched by `grep` and `find`.

A problem: I sometimes play with code and do a lot of plots.

Solution: Use knitr.

An aside: LaTeX and Markdown

```
## An aside: LaTeX and Markdown

Many people use Markdown for

* notes,
* reports,
* documents,
* some math: $\int_0^\infty \sin x/x\, dx = \pi/2$.
```

Markdown: easy to learn, but limited possibilities.

LaTeX: more diffult to learn, but huge possibilities: references, bibliographies, sketches, plots…

Preaching to the choir: LaTeX is a good investment!

knitr

Yihui Xie (2015). *Dynamic Documents with R and knitr*. Second edition. Boca Raton; London; New York: Chapman and Hall/CRC. ISBN: 978-1498716963

A great tool for literate programming and literate science (Boris Veytsman (2014). “Book review: Dynamic Documents with R and knitr, by Yihui Xie”. In: *TUGboat* 35.1, pp. 115–119. URL: `http://tug.org/TUGboat/tb35-1/tb109reviews-xie.pdf`).

knitr example (1)

```
We start from the standard identity
\begin{equation}
  \int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}
\end{equation}
We also add a simple plot
<<device='tikz', fig.width=8, fig.height=3>>=
data <- tibble(x=seq(0.01, 20, by=0.01)) %>%
    mutate(y=sin(x)/x)
ggplot(data) + geom_line(aes(x,y))
@
```

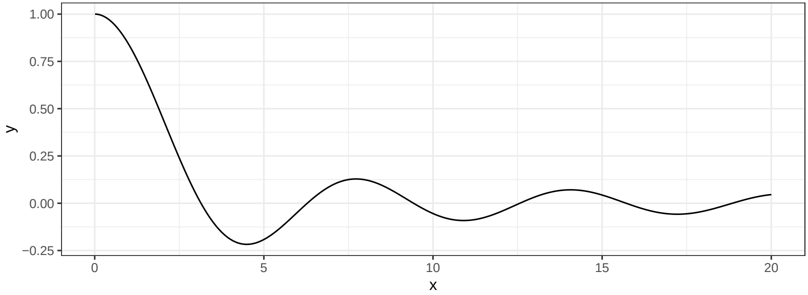
knitr example (2)

We start from the standard identity

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2} \tag{1}$$

We also add a simple plot

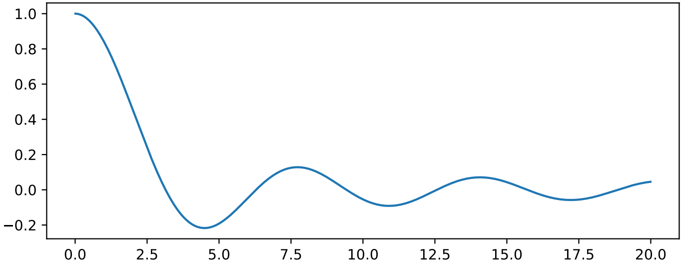
```
data <- tibble(x=seq(0.01, 20, by=0.01)) %>%
    mutate(y=sin(x)/x)
ggplot(data) + geom_line(aes(x,y))
```



Not only R!

Here we use `engine='python'` magic

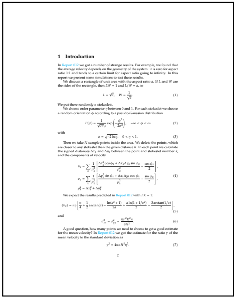
```
import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0.01, 20, 0.01)
y = np.sin(x)/x
plt.plot(x,y)
```

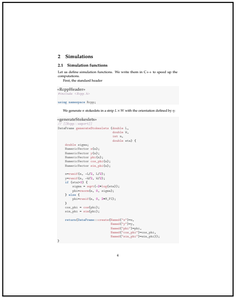


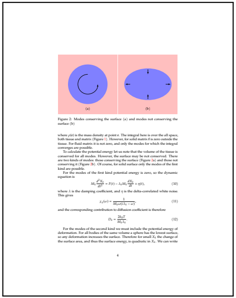
Details, tips and tricks

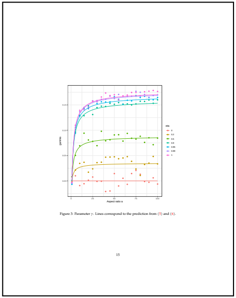
- Start a project with a directory, README and `Makefile` or Rstudio `proj` (or `arara` rules).
- You may need separate directories for data, etc.
- Number notes like `001-introduction.rnw`, `002-hypothesis.rnw`, etc.
- Always use version control!

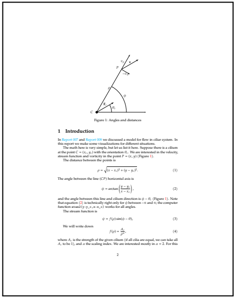
Examples of my lab notes

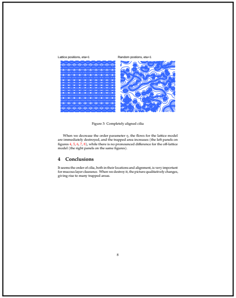












Problems & Solutions

1. Limitations of PDF format: movies & interactive plots are not easy to do! There are solutions, but how reproducible are they? Flash debacle…
2. Speed:
   - I write prose with the speed I think—*good!*
   - I program in knitr with the same speed as in IDE—*good!*
   - I write equations in TeX slightly slower than with a pen—*ok!*
   - I write sketches in TikZ (and in PSTricks) much slower than with a pen—*bad!*.

Solutions for the sketching speed I am considering:

- Doodle with a pen, then scan and use `\includegraphics`.
- Use a program with PDF output.
- Write TikZ faster.

Final exhortation (standing on the shoulders of a giant)

GO FORTH now and create *beautiful, clear and reproducible laboratory notes!*